Clustering-Correcting Codes

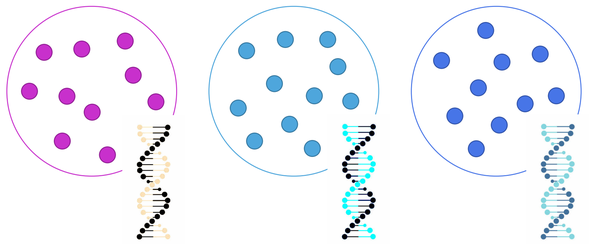
Tal Shinkar, Eitan Yaakobi, Andreas Lenz, Antonia Wachter-Zeh (ISIT 2019)

TECHNION Israel Institute of Technology

Technical University of Munich

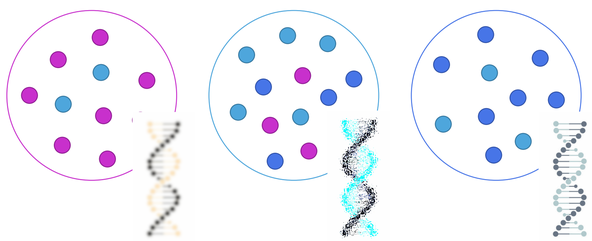
Introduction

Short DNA strands representing some binary data are synthesized and stored as unordered sequences, thousands of times each. Each strand (data) is prefixed with an index so it can be placed in the correct location in respect to the other strands [1].



The Problem

Upon reading each strand has hundreds of copies going through a sequencer. All reads has to be clustered. In the presence of errors in the index part, strands might be misclustered.



The strands are scrambled and it might also affect the recovered data.

Previous Approaches

Several approaches have been offered to solve the clustering problem. Each prioritize computational complexity or storage rate over the other.

· Indices with error-correcting code [2]

- Low storage rate (High redundancy).
+ Efficient clustering (Group by index).

· Clustering by pairwise comparisons [3]

- High computational complexity.
- Does not provide perfect clustering.
+ High storage rate (No redundancy).

Bounds

$\mathcal{A}_{M,L}(\mathrm{e},t)$ denotes the size of the largest CCC. The following bound apply:

$$\frac{\mathcal{A}_{M,L}(e,t)}{2^{ML_M}} \leq \left(1-\frac{B_{L_M}(t-1)}{2^{L_M}}\right)^{M-1}$$

$$\frac{\mathcal{A}_{M,L}(e,t)}{2^{ML_M}} \geq \left(1-\frac{B_{\log M}(e)\cdot B_{L_M}(t-1)}{2^{L_M}}\right)^{M-D_{\log M}(e+1)}$$

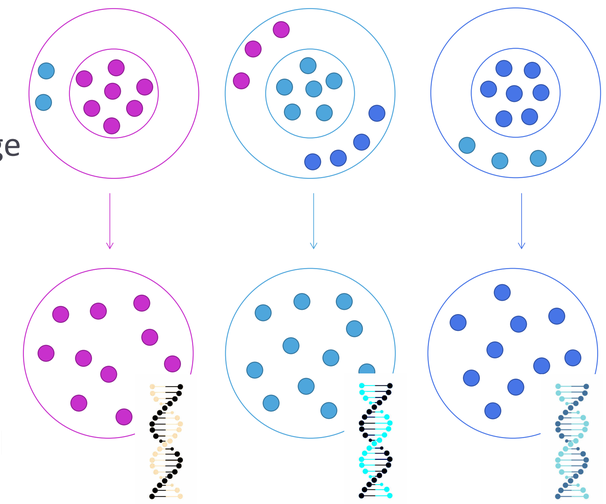
Ensuring Separation

In order to fix a strand that violates the clustering constraint a special sequence is embedded in its data called **Repelling Sequence.** The sequence is built based on all neighbors of the violating strand.

Hybrid Solution

Clustering-Correcting Codes – a new family of constraint codes in which each codeword, when stored into a DNA storage system, can be **efficiently clustered** upon reading. The result clustering is perfect.

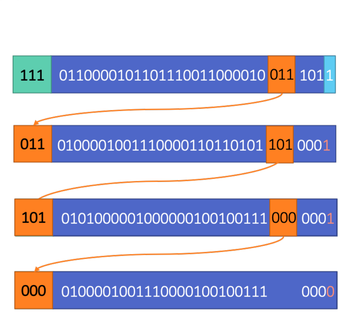
Each codeword satisfies a clustering constraint – for each two indices that can appear as a noisy copy of each other the data parts of the matching strands should be well separated.



Constructing CCCs

Using only 1 bit for redundancy, each group of strands can be encoded efficiently to satisfy the clustering constraint.

Each strand that violates the constraint in respect to some other strand is zipped based on their similarity. The remaining space is used to ensure that their data is well separated as the constraint demands.



All modifications are chained. The decoding is done by going through the chain in reverse order and extracting the original data of each strand.

References

[1] Y.M. Chee, H.M. Kiah, and H. Wei, **"Efficient and explicit balanced primer codes,"** https://arxiv.org/abs/1901.01023, 2019.
[2] M. Blawat, K. Gaedke, I. Hu¨tter, X.M. Chen, B. Turczyk, S. Inverso, B.W. Pruitt, and G.M. Church, **"Forward error correction for DNA data storage,"** Int. Conf. on Computational Science, vol. 80, pp. 1011–1022, 2016.
[3] C. Rashtchian, K. Makarychev, M. Racz, S. Ang, D. Jevdjic, S. Yekhanin, L. Ceze, and K. Strauss, **"Clustering billions of reads for DNA data storage,"** NIPS, 2017.

Computer Science Department, Technion